Journal of Multivariate Analysis 102 (2011) 118–129

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Low dimensional semiparametric estimation in a censored regression model

Ignacio Moral-Arce [a,*], Juan M. Rodríguez-Póo [b], Stefan Sperlich [c]

[a] Instituto de Estudios Fiscales, 28035 Madrid, Spain
[b] Instituto Cantabro de Estadística (ICANE), Santander, Spain
[c] Université de Genève, Département d'économétrie Bd du Pont d'Arve 40, CH-1204 Genève, Switzerland

ARTICLE INFO

Article history:
Received 10 December 2009
Available online 20 August 2010

AMS subject classifications:
62E20
62F12
62G20

Keywords:
Semiparametric censored regression
Partial linear additive models
Marginal integration

ABSTRACT

A new estimation procedure for a partial linear additive model with censored responses is proposed. To this aim, ideas of Lewbel and Linton [A. Lewbel, O. Linton, Nonparametric censored and truncated regression, Econometrica 70 (2002) 765–779] on censored model regression are combined with those of Kim et al. [W. Kim, O. Linton, N.W. Hengartner, A computationally efficient estimator for additive nonparametric regression with bootstrap confidence intervals, Journal of Computational and Graphical Statistics, 8 (1999) 278–297] on marginal integration and those on average derivatives. This allows for dimension reduction, interpretability and — depending on the context — for weights yielding computationally attractive estimates. Asymptotic behavior is provided for all proposed estimators.

1. Introduction

Partial linear additive models have been attracting a lot of attention in semiparametric regression analysis. Within this statistical framework most part of the interest has been devoted to the root-n efficient estimation of the parametric part. For example, [29] proposed a least squares spline estimator of the parametric part. Opsomer and Ruppert [22] studied the statistical properties of the parametric part in a backfitting context. Moral-Arce and Rodríguez-Póo [20] provided an alternative but efficient method for a similar model. Schimek [30] estimated partially linear models based on smoothing splines. For a more recent review see [11].

In this article we propose an extension of these popular models to limited dependent variables, namely censored regression. Even though there exist some algorithms, until now, there is little known about the asymptotic behavior of estimators in these kind of regression models. Burman [3] estimated generalized additive models by splines. Alvarez de Uña and Roca-Pardiñas [2] provided a randomly weighted version of the backfitting algorithm that allows for the nonparametric estimation of the effects of the covariates on the response. Lewbel and Linton [17] introduced an estimator for censored and truncated regression. Lewbel [16] proposed a rather simple semiparametric estimator for the parameters when facing a binary response with a complex index. Furthermore, Chen et al. [4] presented a new estimator that converges at the optimal nonparametric rate with a limiting normal distribution. Das et al. [7] and Rodríguez-Póo et al. [27] presented semiparametric extensions of the so-called Tobit II and III, i.e. models where the truncation and censoring respectively, are determined by simultaneous regression equations. The first authors used series estimators and the second ones kernel methods. Chen and Khan [5] proposed an estimation procedure for a censored regression model where the latent regression function has

* Corresponding author.
E-mail addresses: ignacio.moral@ief.meh.es, moral.arce@gmail.com (I. Moral-Arce), stefan.sperlich@unige.ch (S. Sperlich).

0047-259X/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2010.08.006

a partially linear form based on a conditional quantile restriction. Qin and Jing [26] studied in detail the asymptotics of a censored model with empirical likelihood. Liang and Zhou [18] provided the asymptotic normality of the parametric component in a semiparametric partially linear model with right-censored data. Orbe et al., [23] analyzed the effect of several covariates on a censored response variable with unknown probability distribution. Xia and Härdle [32] introduce an efficient, constructible and practicable estimator of partially linear single-index models with applications to time series.

To summarize, we consider a model that is specified semiparametrically, presents some limited dependent variable and enhances additivity. To our knowledge our paper is the first providing asymptotic theory that copes with all the mentioned issues. Root-n consistency of the estimator of the structural parameters is yielded and also the nonparametric additive components can be estimated at the optimal rate.

The paper is organized as follows. In the next section we will introduce the statistical model and the estimators. Then, we use ideas from [25] to estimate average derivatives, including the parametric part of this model. For calculating the nonparametric additive part we use internalized marginal integration along the ideas of Kim et al. [15]. In Section 3 we provide the corresponding asymptotic theory for these estimators. We discuss several extensions, in particular the special treatment of discrete covariates, in Section 4. Finally, Section 5 presents the conclusions. All proofs are given in the Appendix.

2. Model and estimators

Consider a partial linear additive model with censored response, i.e.

$$Y_i^* = \gamma_0 + \gamma^T X_i + \sum_{j=1}^{d} g_j(T_{ij}) + U_i \quad \text{for } i = 1, \ldots, n, \quad \text{and} \quad Y_i = \begin{cases} Y_i^* & \text{if } Y_i^* > 0, \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where (γ_0, γ^T) are a $1 \times 1+p$ unknown parameter vector, g_j are unknown additive functions, $X_i \in \mathbb{R}^p$, $T_i = (T_{i1}, \ldots, T_{id})^T \in \mathbb{R}^d$ exogenous covariates, U_i the disturbance term, and Y_i^* is the latent endogenous variable. Set $g(T_i) = \gamma_0 + \sum_{j=1}^{d} g_j(T_{ij})$, and let us define the conditional mean of the latent variable

$$s(W_i) = \gamma^T X_i + g(T_i), \quad \text{where } W = (X, T) \tag{2}$$

such that the censored model can be expressed as $Y = \max\{0, s(W) + U\}$.

In the first step, we consider an estimator of the multivariate regression function $s(W)$ by means of the techniques developed in [17]. Given the generic observations $\{Y_i, W_i\}_{i=1}^n$, one would like to estimate the conditional mean $s(w) = E(Y_i^* | W_i = w)$ and its derivatives. The weighted least squares

$$\sum_{i=1}^{n} \left[Y_i^* - c_0 - c_1(W_i - w)\right]^2 K((W_i - w)/h_s), \tag{3}$$

with kernel $K(u)$ and bandwidth h_s is infeasible since the variable Y^* is unobservable. In order to overcome this problem one establishes a set of auxiliary regressions: $r_1(w) = E(Y|W = w)$, $r_2(w) = E[I(Y > 0)|W = w]$, and $q[r_1(w)] = E[I(Y > 0)|r_1(W) = r_1(w)]$.

It can be shown that $s(w)$ and its partial derivatives have the following expressions:

$$s(w) = \lambda - \int_{r_1(w)}^{\lambda} \frac{1}{q(r_1)} \mathrm{d}r_1,$$

$$\frac{\partial s(w)}{\partial w_k} = s_k'(w) = \frac{r_{1k}(w)}{r_2(w)},$$

where $r_{1k}(w) = \frac{\partial r_1(w)}{\partial w_k}$, and $w = (w_1, w_2, \ldots, w_{d+p})^T$.

Let $\hat{r}_1(W)$ be the nonparametric regression of Y on W, $\hat{r}_2(W)$ be a nonparametric regression of $I(Y > 0)$ on W, and $\hat{q}(r_1)$ a nonparametric regression of $I(Y > 0)$ on $\hat{r}_1(W)$. Finally, defining $\hat{\lambda} = \max_{i=1,\ldots,n} \hat{r}_1(W_i)$ one sets

$$\hat{s}(w) = \hat{\lambda} - \int_{\hat{r}_1(w)}^{\hat{\lambda}} \frac{1}{\hat{q}(r_1)} \mathrm{d}r_1. \tag{4}$$

Due to the fact that $\hat{r}_1$, $\hat{r}_2$ and $\hat{\lambda}$ are consistent estimations of the true functions, $\hat{\lambda} - \int_{\hat{r}_1(w)}^{\hat{\lambda}} \hat{q}^{-1}(r_1)\mathrm{d}r_1$ is a consistent estimation of $s(w)$. The same holds for the estimation of the partial derivatives, if wanted. Their estimates can be defined by

$$s_k'(w) = \frac{\hat{r}_{1k}(w)}{\hat{r}_2(w)} \quad \text{where } \hat{r}_{1k}(w) = \frac{\partial \hat{r}_1(w)}{\partial w_k} \tag{5}$$

or, alternatively, obtained directly via local polynomial estimation.

In the second step we decompose $s(w)$. An intuitively attractive procedure would be to study now weighted average derivatives of each covariate, especially as the index of $s(w)$ is fully additive. These are popular e.g. in economics to calculate the returns to scale, and easy to interpret (though depending on the weighting). Let $\bar{\gamma}$ be the vector of these weighted